IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant:	Lin-Shan Lee	Attorney Docket No. LEEL121327
Application No.:	10/612,730	Group Art Unit: ---
Filed:	July 1, 2003	Examiner: ---
Title:	METHOD FOR SPEECH-BASED INFORMATION RETRIEVAL IN MANDARIN CHINESE	

LETTER TRANSMITTING PRIORITY DOCUMENTS

Seattle, Washington 98101

TO THE COMMISSIONER FOR PATENTS:

Enclosed is a certified copy of the following application for which a claim of priority under 35 U.S.C. § 119 has been made:

Country	Application No.	Filed	Title
Taiwan	092107121	March 28, 2003	METHOD FOR SPEECH-BASED INFORMATION RETRIEVAL IN MANDARIN CHINESE

If the Examiner has any questions, please contact the undersigned.

Respectfully submitted,

CHRISTENSEN O'CONNOR
JOHNSON KINDNESS[PLLC]

Shoko I. Leek
Registration No. 43,746
Direct Dial No. 206.695.1780

I hereby certify that this correspondence is being deposited with the U.S. Postal Service in a sealed envelope as first class mail with postage thereon fully prepaid and addressed to Mail Stop - Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the below date.

Date: July 31, 2003

SIL:DDP

LAW OFFICES OF

CHRISTENSEN O'CONNOR JOHNSON KINDNESS[PLLC]
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

中華民國經濟部智慧財產局

INTELLECTUAL PROPERTY OFFICE
MINISTRY OF ECONOMIC AFFAIRS
REPUBLIC OF CHINA

茲證明所附文件，係本局存檔中原申請案的副本，正確無訛，
其申請資料如下 ：

This is to certify that annexed is a true copy from the records of this office of the application as originally filed which is identified hereunder：

申　　請　　日 ： 西元 2003 年 03 月 28 日
Application Date

申　請　案　號 ： 092107121
Application No.

申　　請　　人 ： 李琳山
Applicant(s)

局　　長
Director General

蔡練生

發文日期： 西元 2003 年 7 月 9 日
Issue Date

發文字號： 09220687330
Serial No.

發明專利說明書

（填寫本書件時請先行詳閱申請書後之申請須知，作※記號部分請勿填寫）

※ 申請案號：＿＿＿＿＿＿＿＿ ※IPC 分類：＿＿＿＿＿＿＿＿

※ 申請日期：＿＿＿＿＿＿＿＿

壹、發明名稱

(中文) 以語音為基礎的中文資訊檢索方法

(英文)

貳、發明人（共 4 人）

發明人 1 （如發明人超過一人，請填**說明書發明人續頁**）

姓名：(中文) 李琳山

(英文) LEE LIN-SHAN

住居所地址：(中文) 台北市古亭區大學里 7 鄰溫州街 58 巷 7 號 3 樓

(英文)

國籍：(中文)中華民國 (英文)R.O.C.

參、申請人（共 1 人）

申請人 1 （如申請人超過一人，請填**說明書申請人續頁**）

姓名或名稱：(中文) 李琳山

(英文) LEE LIN-SHAN

住居所或營業所地址：(中文) 台北市古亭區大學里 7 鄰溫州街 58 巷 7 號 3 樓

(英文)

國　籍：(中文)中華民國 (英文)R.O.C.

代表人：(中文) (英文)

發明人 2

姓名：(中文) 簡立峰
(英文)
住居所地址：(中文) 台北縣新店市寶安里 3 鄰中興路 3 段 228 巷 8 號 4 樓
(英文)
國籍：(中文)中華民國 (英文)R.O.C.

發明人 3

姓名：(中文) 陳柏琳
(英文)
住居所地址：(中文) 台北縣汐止市復興里 3 鄰水源路 1 段 80 號 2 樓
(英文)
國籍：(中文)中華民國 (英文)R.O.C.

發明人 4

姓名：(中文) 王新民
(英文)
住居所地址：(中文) 台北縣汐止市白雲里 20 鄰水源路 2 段 22 巷 8 號 10F
(英文)
國籍：(中文)中華民國 (英文)R.O.C.

肆、中文發明摘要

隨著文字、聲音以及多媒體資訊在網際網路上迅速累積並廣泛地被使用，發展以文字或語音型式的查詢指令(text or speech queries)去檢索文字或語音型式的資訊(text or speech information)的技術就顯得愈來愈為重要。以語音為基礎之資訊檢索(speech-based information retrieval)指的是使用者的查詢指令以及被檢索的資訊兩者其中至少之一是語音型式。在本發明中，考慮中文的單音節結構(monosyllabic structure)特性，發展出來一系列以音節(syllable)為基礎的索引特徵(indexing terms)，包括了重疊音節片段(overlapping syllable segments)及可間隔若干音節之雙音節(syllable pairs separated by a few syllables)，同時也驗證了這一系列以音節為基礎的索引特徵的確具有極強的鑑別能力。此外，在本發明裡也發展出進一步融合以中文的字與詞為基礎的索引特徵的方法，以及若干特別的處理方法，來增強上述這些音節索引特徵的檢索鑑別能力。

伍、英文發明摘要

With the rapidly growing use of the text, audio and multi-media information over the Internet, the technology for retrieving text or speech information using text or speech queries is becoming more and more important. By speech-based information retrieval, we mean the user query and/or the information to be retrieved is in the form of speech. In this invention, considering the monosyllabic structure of the Chinese language, a whole class of syllable-based indexing terms, including overlapping segments of syllables and syllable pairs

separated by a few syllables, was developed. The strong discriminating capabilities of such syllable-based indexing terms have been verified. Special approaches for better utilizing such capabilities, including fusion with the word- and character-level information and improved approaches to obtain better syllable-based features and query expressions and so on were developed too.

陸、(一)、本案指定代表圖為：第 1 圖

(二)、本代表圖之元件代表符號簡單說明：

（無）

柒、本案若有化學式時，請揭示最能顯示發明特徵的化學式：

捌、聲明事項

☐ **本案係符合專利法第二十條第一項☐第一款但書或☐第二款但書規定之期間，其日期為：** _____

☐ **本案已向下列國家（地區）申請專利，申請日期及案號資料如下：**

【格式請依：申請國家（地區）；申請日期；申請案號　順序註記】

1. 本案在向中華民國提出申請前未曾向其他國家提出申請專利。
2. _____
3. _____

☐ **主張專利法第二十四條第一項優先權：**

【格式請依：受理國家（地區）；日期；案號　順序註記】

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____

☐ **主張專利法第二十五條之一第一項優先權：**

【格式請依：申請日；申請案號　順序註記】

1. _____
2. _____
3. _____

☐ **主張專利法第二十六條微生物：**

☐國內微生物　【格式請依：寄存機構；日期；號碼　順序註記】

1. _____
2. _____
3. _____

☐國外微生物　【格式請依：寄存國名；機構；日期；號碼　順序註記】

1. _____
2. _____
3. _____

☐熟習該項技術者易於獲得，不須寄存。

(1)

玖、發明說明

(發明說明應敘明：發明所屬之技術領域、先前技術、內容、實施方式及圖式簡單說明)

技術領域：

本發明提供一種資訊檢索方法，尤指一種以語音為基礎的中文資訊檢索方法。

先前技術：

由於網際網路的普及，大量的資訊迅速累積並廣泛地被使用。因此，時空距離遠近不再是人們存取與使用資訊的最大障礙，取而代之的問題是缺乏有效率的方式在浩瀚的網際網路中尋找想要的資訊。資訊檢索技術(information retrieval technologies)因為能夠提供使用者便捷的方式去存取與使用想要的資訊，因此在近幾年來格外地受到重視。直到現在為止，大部分資訊檢索的研究以文字型式的查詢指令(text queries)去檢索文字型式的資訊為主，也就是做文字與文字間的比對，目前在這方面的研究與系統發展已有許多相當不錯的成果。近年來更因為語音辨識技術的進展，開始有一些以整合資訊檢索和語音辨識技術的研究在進行。主要包括了三種不同的應用模式，亦即以語音型式的查詢指令(speech queries)去檢索文字型式的資訊(text information)、以文字型式的查詢指令(text queries)去檢索語音型式的資訊(speech information)和以語音型式的查詢指令(speech queries)去檢索語音型式的資訊(speech information)，上述這三種應用模式我們統稱之為以語音為基礎的資訊檢索(speech-based information retrieval)。值得注意的是，傳統文字型式以外的影音多媒體資訊如廣播、電視節目、數位博

物館等，逐漸大量地出現在網際網路上，顯然已成為文字資訊以外非常重要的資訊來源。在絕大部分的情況下，語音是這些多媒體資訊最主要的組成成分。另一方面，由於輕薄短小的手攜式設備(hand-held devices)像大哥大、PDA等盛行，原本在傳統個人電腦上常使用的輸入裝置如滑鼠、鍵盤等在這些新設備上不是已不復存在，就是不若以往那樣地可以被方便使用，使得語音查詢的功能變得更為受到重視。這些都是為什麼以語音為基礎的資訊檢索變得越來越重要的原因。可以想像在未來這種環境之下，人們可使用手攜式設備以語音查詢指令去檢索多媒體資訊(利用多媒體資訊中的語音組成成分)，將不再是一個可望而不可及的夢想了。當然，有時候使用者的查詢指令或是要被檢索的資訊也可以是文字的形式。對於中文而言，由於中文不是用字母拼成的拼音語言，常用的中文字非常的多，使得中文的電腦輸入即使在今天也一直是一個非常困難而且尚未完全解決的問題。因此，對於中文來說，發展以語音為基礎的資訊檢索技術將會比其他語言來得重要而且更具吸引力。

與傳統文字型式的資訊檢索不同的是，以語音為基礎的資訊檢索並不能直接地拿輸入的查詢指令(queries)來與資料庫中很多筆資訊記錄(information records)一一來作比對。有很多筆資訊記錄和輸入的查詢指令在題旨上可能是相關的，但是由於輸入的查詢指令(queries)與每一筆資訊記錄(information records)彼此的用字遣詞可能不同，或者是聲學

環境(acoustic conditions)、語者(speakers)、講話的模式(speaking modes)和背景雜訊(background noises)等的不同，使得處理上變得更加的困難。因此對於查詢指令與資訊記錄而言，不管它們是以文字或是語音的形式存在，都必須先適當地轉換成某種代表資訊內涵的索引特徵(indexing terms)以用來判斷查詢指令與資訊記錄之間的相關程度。因此，如何在詞彙、主題與聲學環境都充滿不確定變異性的情況下能正確辨識中文語音進而從事語音資訊檢索，就是首要問題之所在。這些變異因素使得完全正確的語音辨識不可能達成，反而不可避免地產生一定程度的錯誤辨識結果。而為了克服這些錯誤的辨識結果所造成的影響，當然會使得本發明所提出的以語音為基礎的資訊檢索技術與傳統的文字型式的資訊檢索(所有文字都是正確的)截然不同，而必須要具備了相當程度的強健性(robustness)才可以。

中文的以語音為基礎的資訊檢索第二個主要問題，便是要選擇適當的索引特徵(indexing terms)來同時描述使用者查詢指令及所要查詢的每一筆資訊記錄，使得它們彼此間的相關性在檢索過程中可以很容易地被評估出來。索引特徵的選擇主要有兩種作法：一種是僅以關鍵詞(Keyword)作為索引基礎(keyword-based approach)，另一種則是以所有的詞彙作為索引基礎(word-based approach)。對於前者僅以關鍵詞作為索引基礎的方法，必須事先為要被檢索的每一筆資訊記錄定義好一組關鍵詞(keywords)，再從使用者輸入的查詢指令中擷取出可能的關鍵詞，這樣一來，含有與查詢指令相

同或相關的關鍵詞之資訊記錄就可以檢索出來。這種方法非常簡易，尤其是對於檢索相對靜態(static)的資訊記錄，因為主要可供搜尋的關鍵詞並不會經常改變。然而就算事前已經知道了要被檢索的資訊記錄的內容，如何為它們定義一組完善的關鍵詞組卻並不是一件非常容易的事。尤其在網際網路的環境下，資訊記錄是每天持續不斷在累積改變並非全然靜態的，使用先前定義好的關鍵詞組幾乎不可能滿足這樣的檢索需求，不管定義的關鍵詞組多大，遺漏關鍵詞的情況總是一定會發生。有了這一考量後，很自然地會想到以所有的詞彙當為索引的作法。當使用者查詢指令與所有的資訊記錄都被完整的以文字表示後(可能以中文的字或詞的方式呈現，查詢指令與資訊記錄兩者都可以是經由語音辨識技術產生的)，許多已發展很好的文字型式的資訊檢索技術就可以直接地使用。然而，即使是採用這種以所有的詞彙當作索引特徵的作法，詞典外詞彙(Out-of-vocabulary，亦即用了不少語音辨識器的詞典中所沒有的詞，語音辨識器一定辨識不出來)的發生仍會是一個問題。因為大詞彙語音辨識器中通常需要一個事先定義好的詞典，但有些對於資訊檢索而言是特別重要的關鍵詞，可能因為沒有被包括在這個詞典裡而沒有辦法被辨識出，這對中文來說是確實存在的問題，將在下一節詳細說明這個問題。這個問題因而引出直接在比"詞"更小的層次上比對查詢指令及資訊記錄的相關性的概念。因為，在這種情況下，並不一定需要有"詞"這一層次，語音資訊

檢索也就不會受限於語音辨識辭典大小的影響。

發明內容：

在本發明中，考慮中文單音節結構(monosyllabic structure)特性，發展出一系列以音節(syllable)的統計特性為基礎的索引特徵(indexing terms)來從事中文的以語音為基礎的資訊檢索，並驗證了這一系列以音節為基礎的索引特徵在檢索表現上的確具有極強的鑑別能力。同時，也進一步融合了以中文的字與詞為基礎的索引特徵並發展出若干特別的處理方法來增強上述這些索引特徵在檢索上的表現。

實施方式：

I. 使用音節層次統計特性的理由

在中文裡是一字一音，每個字(至少有一萬個以上的常用字)都是發一個單音節(monosyllable)的音。中文有一大特色，即是每天都很容易有新詞(new words)產生，新詞通常是由一到數個字或者音節結合而成。例如，由“電”與“腦”這兩個字的結合形成了一個新詞“電腦”，由“股”、“市”、“長”和“紅”這四個字結合可形成了一個金融交易上的新詞“股市長紅”。在大部分的情況下，這些新詞的語意多少都跟構成它們的字有關。另外許多專有名詞像人名、組織名、地名及特定領域的術語等也都是會隨時產生的新詞，就如同上面所舉的例子一樣。因為這些新詞都帶有特定的關鍵的語意足以描述或分辨不同的主題對象，所以在資訊檢索功能上往往都是常會被用來查詢的重要詞彙。但是在絕大多數的情況下，這些對於檢索

而言非常重要的詞彙卻常常完全沒有包含在語音辨識器的詞典裡。因此在從事以語音為基礎的中文資訊檢索時，詞典外詞彙(out-of-vocabulary)發生的情況特別的嚴重，這也就是為什麼本發明以音節層次的統計特性(syllable-level statistical characteristics)的索引特徵來解決這些在資訊檢索常發生的問題是有道理的。換句話來說，在中文裡適當的音節組合可以代表發相同音之對應字組合的語意，而這些音節組合來當作索引特徵，就可避免資訊檢索時需以詞當作索引特徵時會遭遇的詞典外詞彙問題。

事實上，中文具有獨特的一字一音節的發音結構，使得以音節層次資訊(syllable-level information)來從事以語音為基礎的中文資訊檢索，的確有其非常重要的意義。雖說中文的常用字至少有一萬個以上，但由於中文獨特的一字一音節結構特性，以及許多截然不同語意的字可對應到同一個音節，使得中文的音節數目僅有 1,345 個。由於每個詞是由一到數個字(或音節)所組合而成，於是這 1,345 個音節就可以組合成無限多個中文的詞。也就是說，雖說每個音節是對應到許多含不同語意的字，然而由數個特定的音節組合在一起卻常僅產生唯一的多音節詞(polysyllabic words)，或偶而有極少的同音多音節詞(如程式、城市)。因此，若在從事資訊檢索時以數個音節組合成的片段為特徵來比較輸入的查詢指令與被檢索的資訊記錄將可以提供非常好的檢索評估依據。

另一方面，採用音節層次資訊(syllable-level information)來

從事資訊檢索其實還存在有許多的重要原因。在中文裡，幾乎每個字都是一個本身具有語意的詞素(morpheme)，在語言上可以有相當獨立的角色。所以，由數個字構成詞時，構詞往往非常的有彈性。舉例來說，在多數的情況下，描述相同或相似概念的詞可能僅有其中的一兩個字是不同的，其餘的字都是相同的。譬如“中華文化”和“中國文化”是描述相同的語意，但是它們的第二個字是不同的。另一個可以觀察到的現象是在中文裡，一個長詞可以隨意地縮寫成較短的詞，譬如保留“國家科學委員會”的第一個、第三個以及最後一個字就可以縮寫成“國科會”。再者，時常一個由外國語言引入的詞(exotic word)根據它的發音可以翻譯成不同的詞，例如“Kosovo”可以翻譯成“科索沃/ke1-suo3-wo4/”、“柯索佛/ke1-suo3-fo2/”、“克索夫/ke1-suo3-fu1/”、“科索伏/ke1-suo3-fu2/”、“科索佛/ke1-suo3-fo2/”等等，但這些經翻譯過的詞通常都含有一些音節是或者全部的音節都是相同的。為此，一個智慧型的檢索系統必須要能夠處理中文彈性的構詞現象，當查詢指令與被檢索的資訊記錄有不同的詞卻描述近似的語意時，相關的資訊記錄還是可以被成功地檢索出來。直接在音節層次比對語音查詢指令與語音資訊紀錄的相關性的確可以在某種程度上解決上述中文彈性構詞問題，因為在檢索的過程中“詞”並不一定需要被辨識出來，而且不同形式的詞若是描述相同或相關概念，常都含有一些相同的音節。

II. 核心技術

A. 音節層次索引特徵 (Syllable-level Indexing Terms)

本發明提供了一系列以音節(syllable)為基礎的索引特徵，包括了以不同長度的重疊音節片段(overlapping syllable segments with length N, S(N), N=1,2,3,4,5, etc.)及間隔若干音節之雙音節(syllable pairs separated by a few syllables , PS(n), n=1,2,3,4, etc.)為索引特徵的技術。以一個長度為 10 的音節序列(a syllable sequence of 10 syllables S_1 S_2 S_3.... S_{10})為例，前者(不同長度的重疊音節片段)列在圖一的上半部，後者(間隔若干音節之雙音節)則列於圖一的下半部。例如長度為 3 的重疊音節片段(S(N), N=3)包括了音節片段(S_1 S_2 S_3)、(S_2 S_3 S_4)、(S_3 S_4 S_5)等等，間隔一個音節之雙音節($P_S(n)$, n=1)有(S_1 S_3), (S_2 S_4), (S_3 S_5)等等。考慮中文語言的結構性特徵，上述這些音節層次的索引特徵的確是在檢索過程中是有意義的。如同上面所提及的，每一個音節其實代表(對應)許多不同語意的字，而且若兩個詞代表相似或相關的概念，經常它們的組成音節中有一些是相同的，即使當中有的詞是屬於詞典外詞彙(out-of-vocabulary)，語音辨識器無法辨識出來。因此以長度為 1 的音節片段(S(N), N=1)來作為索引單位，在檢索上是有其道理的。然而，由於每一個音節同時對應到許多代表不同語意的同音字，如果僅用長度為 1 的音節片段(S(N), N=1)來作索引，在檢索時必定會發生嚴重的混淆問題，因此必須要再結合其他的索引特徵才行。事實上，在中文 5,000 個最常用的多音節詞裡(polysyllabic words)約百分

之九十以上的詞是雙音節詞，也就是說它們是發兩個音節的音。所以，以長度為 2 的音節片段(S(N), N=2)來作為索引特徵絕對會保有大多數語言上的資訊，在檢索上成為重要索引特徵是有其道理的。同樣地，如果長度較長的音節片段如長度為 3 的音節片段(S(N), N=3)在檢索比對時同時出現在查詢指令與被檢索的資訊記錄中時，與查詢指令有關的重要資訊便可以更精確地被擷取出。另一方面，就上述中文構詞之彈性而言，以間隔若干音節之雙音節來當作索引特徵在檢索上是會有幫助的。就以前述所舉的例子來說，“國家科學委員會”這個詞可以被縮寫或唸成“國科會”，僅包括了原來的第一個、第三個以及最後一個音節，因此本發明所提出的以間隔若干音節之雙音節(syllable pairs separated by n syllables)為索引的方法就明顯地可以解決這個問題。再者，由於在中文語音辨識過程中常有音節的取代(substitution，亦即一個音節被辨識成另一個音節)、插入(insertion，亦即在兩個相連的音節中間，辨識的結果會多出一個不存在的音節)以及刪除(deletion，亦即一個明明存在的音節在辨識時被丟掉)等錯誤的發生，本發明所提出的以間隔若干音節之雙音節為索引(syllable pairs separated by n syllables)的方法也同樣地可以降低這些語音辨識錯誤在檢索上的影響。總而言之，單音節(monosyllables)所形成的索引特徵其實代表著某些具有語意的字，也可以或多或少地解決中文的詞典外詞彙的問題。而不同語意的同音字對應到相同音節所產生的混淆問題，也可以由長度大於 1

的重疊音節片段(overlapping syllable segments with length N, N>1)以及間隔若干音節之雙音節(syllable pairs separated by n syllables)所形成的索引特徵來區分出不同的語意資訊。重疊音節片段為索引特徵可以代表多音節詞或片語(polysyllabic words or phrases)的資訊，對於檢索來說是非常重要的；間隔若干音節之雙音節為索引單位可以在某種程度上解決中文彈性構詞問題如縮寫等，以及降低語音辨識產生的取代(substitution)、插入(insertion)以及刪除(deletion)等錯誤所造成的影響。

當定義好上述一系列以音節(syllable)為基礎的索引特徵後，對於每一項語音查詢指令與每一筆語音記錄都經語音辨識產生對應的音節格狀組(syllable-lattice)。在這音節格狀組中，每個一個音節的語音段落，都儲存著許多的候選音節(syllable candidates)，這是為了克服語音辨識的不確定性，多保留一些候選音節可以確保正確音節沒有流失。同時，每個一個候選音節都存有經語音辨識過程產生的聲學辨識分數，而對於上述的每一音節組合所形成的索引特徵，索引特徵的分數就是由它們個別的組成音節的聲學辨識分數平均而得。若查詢指令或資訊記錄中的任一者是文字型式，則該索引特徵的分數就由其在文字型式的查詢指令或文字型式的資訊記錄中出現的次數來替代。

有了本發明的一系列以音節(syllable)為基礎的索引特徵用來描述語音查詢指令與每一筆語音記錄，則目前許多常在文字型式的資訊檢索(text-based information retrieval)系統使

用的資訊檢索模型(information retrieval models)也都同樣地可以拿來用在以語音為基礎之資訊檢索中使用。就以最常用在文字型式的資訊檢索的向量空間模型(vector space model, 這是所有做文字型式的資訊檢索的人都熟知的技術)來說，在這個模型下，不論資訊紀錄與查詢指令是文字型式或是語音型式，都可以設計一組特徵向量來描述它們，其中的每一個向量分量(component)代表某一類以音節為基礎的索引特徵在檢索時對應的資訊。舉例來說，若使用本發明所提出的各類音節層次的索引特徵中的 9 類加以組合(S(N), N=1~5, and Ps(n), n=1~4)，就一共可以用 9 個特徵向量來代表每一筆資訊紀錄與每一項查詢指令。而資訊紀錄與查詢指令間的相關性就以查詢指令及每一筆資訊記錄的這 9 個特徵向量的個別比對結果的加權和來評估，就似傳統文字型式的資訊檢索的處理過程是完全一樣的。

B. 音節、字與詞三個層次的資訊的融合(Fusion of Syllable-, Character- And Word-Level Information)

雖然上述以音節組合為基礎的索引特徵已經可以在以語音為基礎的中文資訊檢索(speech-based information retrieval for Mandarin Chinese)中提供非常強的鑑別能力，字與詞層次上的資訊卻也可以帶來不少音節所沒有的額外知識。例如，同音字對應到相同音節所衍生的混淆問題可由字層次上的資訊來解決，詞則具有較音節更為完整的語意資訊。但另一方面，以字或詞組合為索引特徵在以語音為基礎的資訊檢索中會帶有較多的語音辨識錯誤，尤其是因詞典外

詞彙引起的辨識錯誤。因此適當地融合音節、字與詞這三種不同層次的資訊，自然就會對於以語音為基礎的中文資訊檢索會有所幫助。就如同前述的音節層次的索引特徵，字與詞層次的索引特徵也可以經由同樣的方式產生，譬如不同長度的重疊字片段或重疊詞片段(C(N), N=1,2,3,4,5, etc., and W(N), N=1,2,3,4,5, etc.)和間隔若干字或詞之雙字或雙詞($P_C(N)$, N=1,2,3,4, etc., and $P_W(N)$, N=1,2,3,4, etc.)。如此一來，查詢指令與資訊記錄間的相關程度就可以用上述音節、字與詞這三種層次的索引特徵個別的特徵向量相關性比對結果的加權和來評估。

C.由資料庫導引的索引特徵(Data-Driven Indexing Terms)

上述以不同長度的重疊音節片段(overlapping syllable segments with length N, S(N), N=1,2,3,4,5, etc.)，字片段或詞片段為索引特徵的方式，效果雖好，但因這些索引特徵的總數龐大，對計算量及記憶體容量的需求極大，實際製作時之軟硬體代價較高。改進的方法，可以進一步利用統計的方法，用電腦程式自動地從資料庫(例如所有被檢索的資訊記錄所形成之集合等)中尋找結合性強且語意完整的音節片段(或字片段、詞片段)為真正使用的索引特徵，而把語意不完整的音節片段(或字片段、詞片段)全部刪除。例如音節片段或字片段“柬埔寨/jian3-pu3-zhai4/”(S(N)或 C(N), N=3)會被選為真正使用的索引特徵，而音節片段或字片段“柬埔/jian3-pu3/”及“埔寨/pu3-zhai4/”(S(N)或 C(N), N=2)等則會因語意不完整，自動地被刪除。這種由資料庫

導引(data-driven)概念下所挑選出來的索引特徵，不僅可以達到非常精簡的索引特徵總數，而且其檢索的效能也會大幅地提高。此概念相同適用於音節、字及詞三個層次的索引特徵。以詞片段舉例，"陳水扁總統"是"陳水扁"及"總統"二個詞所構成的語意完整的雙詞片段，是很好的索引特徵，但"總統前往"是"總統"和"前往"兩個詞，但連起來其語意並不完整，不是一個很有意義的雙詞片段，在檢索時實際意義不大，則可刪除。這種由資料庫導引的索引特徵(data-driven indexing terms)之產生方法，以音節層次的索引特徵為例，可由全體長度為 1 的音節片段(S(N), N=1)開始，以由下而上(bottom-up)的方式，選定結合性強，適於結合的相連音節片段，一一予以兩兩相連形成長度較大的(N=2,3 等)新的音節片段，結合的依據取決於任意兩個在資料庫(例如所有被檢索的資訊記錄所構成的集合等)中相連的音節片段在整個資料庫中的某些統計數值，例如他們彼此間的相互訊息量(mutual information)及語言模型參數(language model parameter)等相當程度代表其結合性的統計數值，或其他類似的統計數值，再對不同長度的索引特徵給予不同的閾值 d_0 設定。當兩個相連的音節片段的某些統計數值大於閾值 d_0 時，便可把他們結合在一起以形成新的音節片段。此一產生步驟可用電腦程式反覆進行若干次，直到沒有任何相連的音節片段的這些統計數值超過閾值為止。同樣的方法也適用於產生由資料庫導引的結合性強且語意完整的字片段或詞片段等等。

D. 音節層次的聲音確認 (Syllable-level Utterance Verification)

當在音節格狀組 (syllable-lattice) 中，每個一個音節的語音段落所儲存的候選音節數目由 1 增加到 m 時，則重疊音節片段 (overlapping syllable segments with length N, S(N), N=1,2,3,4,5, etc.) 及間隔若干音節之雙音節 (syllable pairs separated by a few syllables , $P_S(n)$, n=1,2,3,4, etc.) 的索引特徵數目就會分別增加到 m^N 與 m^2 倍之多。雖說它們之中可能會有一個重疊音節片段或者間隔若干音節之雙音節會是完全正確並因此可以提供適當的檢索資訊，但其餘的 m^N-1 或 m^2-1 個索引特徵都包含有一個以上的錯誤音節，因此不可避免地產生錯誤的索引特徵造成檢索過程中的干擾。音節層次的聲音確認技術於是可以在這裡使用，以降低錯誤索引組合的數目。基本的作法是任何候選音節若其聲學辨識分數低於某個事先設定的閾值 (pre-assigned threshold) 時，其產生的索引特徵就可以被刪除。可以在建立索引特徵時，對每一類索引特徵給不同的閾值的設定。

E. 低頻索引特徵的刪除 (Deletion of Low Frequency Indexing Terms)

可以假設語音辨識結果中含有出現頻率較低的音節組合之處經常較有可能含有辨識錯誤，所以在索引特徵產生過程中，某一索引特徵若含有極低頻率的音節組合成分時，便可予以刪除。因此在本發明中，索引特徵的統計分佈可以用來作為另一種索引特徵刪減的依據。上述的重疊音節片段 (overlapping syllable segments with length N, (S(N),

N=1,2,3,4,5, etc.)及間隔若干音節之雙音節(syllable pairs separated by a few syllables , $P_S(n)$, n=1,2,3,4, etc.)等每一索引特徵的統計分佈，便可以用來作為索引特徵刪減的依據。就舉長度為 2 的重疊音節片段(S(N), N=2)為例，若一個由兩個音節組合成的音節片段(s_k,s_j)其出現次數小於一個事先決定的閾值 r_0 時，便可刪除它以增進檢索的效能。同樣地，對每一類索引特徵可以給不同的閾值的設定。

F. 極高頻索引特徵的刪除(Deletion of Stop Terms)

當產生音節、字與詞的索引特徵時，可針對個別索引特徵的文件倒數頻率(Inverse Document Frequency, IDF，這是一般文字型式的資訊檢索常用的參數)或其他類似的參數為基礎，建立極高頻索引特徵列表(stop term list)。這些是最不具鑑別能力的索引特徵。例如"的　"　"是"這兩個單字或單音節大量出現在每一筆資訊記錄中，故完全沒有索引功能。因此對於每一類音節索引特徵，例如重疊音節片段(overlapping syllable segments with length N, S(N), N=1~5)及間隔若干音節之雙音節(syllable pairs separated by a few syllables , S(N), N=1~5)等，都可建立一個極高頻索引列表，並在產生索引特徵時把每一類索引特徵裡出現在極高頻索引列表中的前 M 個最常出現的索引特徵(亦即 IDF 值較低者等等)從特徵向量中刪除。這裡M的值亦可以依每一類索引特徵而設定。

G. 自動相關迴授(Automatic Relevance Feedback)

在檢索的過程中使用者往往未必能一句話就說出最正

確的查詢指令，有時某些對檢索目的而言是極重要的檢索的線索的索引特徵並沒有出現在使用者的查詢指令中，導致在第一次檢索時並不一定能完全檢索到想要的資訊紀錄。此時，在第一次檢索時找到的相關或不相關資訊記錄(relevant or irrelevant information records)可以用來自動進行第二次檢索，進一步確認使用者實際上真正想要尋找的資訊為何。自動化相關迴授就是把第一次檢索到的，可能是使用者想要的相關資訊記錄中常出現的索引特徵加入使用者的初始查詢指令的特徵向量中，或將在第一次檢索中認為不相關的資訊記錄中常出現的索引特徵從使用者的初始查詢指令的特徵向量中刪除，再以所產生的新的查詢指令特徵向量來從事第二次的檢索，通常均可增進檢索的準確性。

H. 索引特徵關連矩陣 (Term Association Matrix)

如果兩個索引特徵常常同時出現(co-occurring)在相同的資訊紀錄或段落(information records or passages)中，往往可能是共同用來描述某個特定的事件、領域或主題的，因此彼此之間可能存在某種程度上的同義關連性(synonymity association)。基於這樣的假設，可以從要被檢索的資訊記錄所形成的集合中，為每一類的索引特徵建立起一個索引特徵關連矩陣，在此關連矩陣中每一個元素 a(m,n)代表著任兩個索引特徵 t_m 和 t_n 同時出現在相同資訊紀錄或段落的頻率統計特性，因此也代表著這兩個索引特徵之間的某種關連性。例如，若關連矩陣中某一個元素 a(m,n)的值為 1，

可能代表著索引特徵 t_m 和 t_n 總是同時出現在相同的資訊紀錄或段落中，因此一定有非常高的同義關連性；若關連矩陣中某一個元素 a(m,n)的值為 0，可能代表著索引特徵 t_m 和 t_n 從來沒有同時出現在相同的資訊紀錄或段落中，故可能是毫無關係的。於是，我們便可以把與使用者的初始查詢指令中的索引特徵的同義關連性最大的 L 個索引特徵加入查詢指令的特徵向量中，以形成新的使用者查詢指令特徵向量。L 值的大小，可因不同類的索引特徵而異。

實施例流程圖

請參照圖 2，圖 2 為本發明一實施例之流程圖。其中結合了上述以不同長度的重疊音節/字/詞片段或相隔若干音節/字/詞之雙音節/字/詞的方法進行檢索、由資料庫導引的索引特徵抽取方法、經由音節層次的聲音確認、索引特徵關連矩陣、低頻索引刪除與極高頻索引刪除、同時融合音節/字/詞索引特徵及自動相關迴授完成本發明之以語音為基礎之中文資訊檢索。

以上所述僅為本發明之較佳實施例，凡本發明申請專利範圍所做之均等變化與修飾，皆應屬本發明之涵蓋範圍。

圖式簡單說明：

圖 1 為以音節序列 S_1 S_2 S_3..... S_{10} 為例的各種音節層次的索引特徵示意圖。

圖 2 為本發明一實施例之流程圖。

拾、申請專利範圍

1. 一種中文資訊檢索方法，包含：

輸入描述所欲查詢資訊之語音或文字查詢指令；

決定一種索引特徵；及

利用該索引特徵檢索所欲查詢之以語音或文字型式呈現之資訊記錄，

其中該索引特徵係為具有一特定長度的重疊音節片段，且該特定長度可任意指定且至少為一。

2. 如申請專利範圍第 1 項之中文資訊檢索方法，其中該特定長度係為二。

3. 如申請專利範圍第 1 項之中文資訊檢索方法，其中該特定長度係為三。

4. 一種以語音為基礎的中文資訊檢索方法，包含：

輸入描述所欲查詢資訊之語音或文字查詢指令；

決定一種索引特徵；及

利用該索引特徵檢索所欲查詢之以語音或文字型式呈現之資訊記錄，

其中該索引特徵係為一間隔至少一音節之雙音節。

5. 如申請專利範圍第 1 項之中文資訊檢索方法，其中該索引特徵亦可為具有一特定長度的重疊字片段，且該特定長度可任意指定且至少為一。

6. 如申請專利範圍第 1 項之中文資訊檢索方法，其中該索引特徵亦可為具有一特定長度的重疊詞片段，且該特定長度可任意指定且至少為一。

7. 如申請專利範圍第4項之中文資訊檢索方法，其中該索引特徵亦可為一間隔若干字之雙字。

8. 如申請專利範圍第4項之中文資訊檢索方法，其中該索引特徵亦可為一間隔若干詞之雙詞。

9. 如申請專利範圍第1、4、5、6、7或8項之中文資訊檢索方法，其中該索引特徵可經選定為不只一種。

10. 如申請專利範圍第1、4、5、6、7或8項之中文資訊檢索方法，其中該索引特徵可由重疊音節片段、雙音節、重疊字片段、重疊詞片段、雙字及雙詞所組成之群組中選定一或多種。

11. 如申請專利範圍第1、4、5、6、7或8項之中文資訊檢索方法，其中該索引特徵決定後，該中文資訊檢索方法另包含：

辨識語音查詢指令中每一音節、字或詞之語音段落產生一個或一個以上候選音節、字或詞，以建立對應之音節、字或詞格狀組；及

辨識語音資訊記錄中每一音節、字或詞之語音段落產生一個或一個以上候選音節、字或詞，以產生對應之音節、字或詞格狀組；其中該音節、字或詞格狀組中之各候選音節、字或詞包含有經語音辨識產生之一聲學辨識分數。

12. 如申請專利範圍第11項之中文資訊檢索方法，其中該索引特徵另包含有一分數，且該分數係由該索引特徵所包含之所有候選音節、字或詞之聲學辨識分數平均而

得。

13. 如申請專利範圍第1、4、5、6、7或8項之中文資訊檢索方法，其中以語音為基礎之中文資訊檢索係包含有以語音型式的查詢指令檢索文字形式的資訊記錄、以文字型式的查詢指令檢索語音形式的資訊記錄、及以語音形式的查詢指令檢索語音形式的資訊記錄。

14. 如申請專利範圍第13項之中文資訊檢索方法，其中查詢指令或資訊記錄凡以文字型式呈現者，其索引特徵的分數係為該索引特徵在該文字形式的查詢指令或資訊記錄中出現的次數。

15. 如申請專利範圍第1、4、5、6、7或8項之中文資訊檢索方法，另包含為每一查詢指令及每一資訊記錄設計一組特徵向量，其中每一特徵向量包含有若干個向量分量，每一向量分量係用以代表前述中文資訊檢索中每一索引特徵在查詢指令與資訊記錄中由聲學辨識分數求得的分數(若為語音型式呈現)或出現的次數(若為文字型式呈現)。

16. 如申請專利範圍第15項之中文資訊檢索方法，其中該查詢指令與每一資訊記錄之關連性係由代表該查詢指令與代表每一資訊記錄之各特徵向量之個別比對結果的加權和決定。

17. 如申請專利範圍第1、4、5、6、7或8項之中文資訊檢索方法，另包含有產生一組由資料庫導引之索引特

徵，該組索引特徵可由長度為 1 的音節、字或詞片段開始，以由下往上的方式，將相鄰的音節、字或詞片段兩兩相連以形成另一長度較長的音節、字或詞片段，並以該長度較長之音節、字或詞片段在一資料庫中之一統計數值，來決定是否應將該兩音節、字或詞片段加以結合以形成新的索引特徵。

18. 如申請專利範圍第 17 項之中文資訊檢索方法，其中該另一長度較長的音節、字或詞片段之長度為 2。

19. 如申請專利範圍第 17 項之中文資訊檢索方法，其中該另一長度較長的音節、字或詞之長度為 3。

20. 如申請專利範圍第 17 項之中文資訊檢索方法，其中該統計數值可為該可以相連形成另一長度較長的音節、字或詞片段的兩個較小音節、字或詞片段彼此間的相互訊息量。

21. 如申請專利範圍第 17 項之中文資訊檢索方法，其中該統計數值可為該可以相連形成另一長度較長的音節、字或詞片段的兩個較小音節、字或詞片段彼此間的語言模型參數。

22. 如申請專利範圍第 17 項之中文資訊檢索方法，其中該產生由資料庫導引之索引特徵之步驟中，決定是否結合兩個相連的較小音節、字或詞片段以形成另一長度較長的音節、字或詞片段以作為新的索引特徵時，係對不同長度的音節、字或詞片段索引特徵給予不同的閥值，當該統計數值大於該閥值時，便將該兩較小

音節、字或詞片段結合以形成新的索引特徵。

23. 如申請專利範圍第 22 項之中文資訊檢索方法，其中該產生由資料庫導引之索引特徵之步驟可反覆執行，直到沒有任何相連的音節、字或詞片段的統計數值超過該閾值為止。

24. 如申請專利範圍第 11 項之中文資訊檢索方法，其中各候選音節、字或詞之聲學辨識分數若低於一預先設定的值時，該候選音節、字或詞便會被刪除。

25. 如申請專利範圍第 12 項之中文資訊檢索方法，其中該索引特徵在一資料庫中出現之次數若低於一預先設定的值時，該索引特徵便會被刪除。

26. 如申請專利範圍第 25 項之中文資訊檢索方法，其中該預先設定的值可於決定該索引特徵時便加以設定，且不同的索引特徵可設定不同的值。

27. 如申請專利範圍第 1、4、5、6、7 或 8 項之中文資訊檢索方法，另包含有根據各索引特徵之文件倒數頻率建立一極高頻索引特徵列表。

28. 如申請專利範圍第 27 項之中文資訊檢索方法，另包含有從特徵向量中刪除出現在該極高頻索引特徵列表中的前若干個最常出現的索引特徵。

29. 如申請專利範圍第 1、4、5、6、7 或 8 項之中文資訊檢索方法，另包含有為該組索引特徵建立一索引特徵關連矩陣，該矩陣包含若干個矩陣元素，每一矩陣元素代表任兩個索引特徵同時出現在相同的資訊記錄

中的頻率統計特性。

30. 如申請專利範圍第 29 項之中文資訊檢索方法，其中該元素可為介於 0 與 1 之間之任何數值。

31. 如申請專利範圍第 30 項之中文資訊檢索方法，其中該元素為 0 可代表兩個索引特徵從未同時出現在相同的資訊記錄中或無關連性。

32. 如申請專利範圍第 30 項之中文資訊檢索方法，其中該元素為 1 可代表兩個索引特徵總是同時出現在相同資訊記錄中或有非常高的關連性。

33. 如申請專利範圍第 32 項之中文資訊檢索方法，另包含將最具有關連性的若干個索引特徵加入查詢指令的特徵向量中，以形成另一新的查詢指令特徵向量。

34. 如申請專利範圍第 1、4、5、6、7、8、12 或 14 項之中文資訊檢索方法，另包含有於利用該索引特徵檢索欲查詢之以語音或文字型式呈現之資訊記錄之步驟後，進行一第二次檢索。

35. 如申請專利範圍第 34 項之中文資訊檢索方法，其中該第二次檢索可由增加索引特徵或刪除索引特徵，以產生另一新的查詢指令特徵向量加以執行。

36. 如申請專利範圍第 35 項之中文資訊檢索方法，其中該索引特徵之增加或刪除可由該索引特徵常出現於之前檢索所獲得之相關資訊記錄或不相關資訊記錄中加以判斷。

37. 如申請專利範圍第 36 項之中文資訊檢索方法，其中

若該索引特徵常出現於之前檢索所獲得之相關資訊記錄中，則增加該索引特徵或其分數。

38. 如申請專利範圍第 36 項之中文資訊檢索方法，其中若該索引特徵常出現於之前檢索所獲得之不相關資訊記錄中，則刪除該索引特徵或降低其分數。

39. 如申請專利範圍第 11 項之中文資訊檢索方法，另包含有於利用該索引特徵檢索欲查詢之以語音或文字型式呈現之資訊記錄之步驟後，進行一第二次檢索。

40. 如申請專利範圍第 39 項之中文資訊檢索方法，其中該第二次檢索可由增加索引特徵或刪除索引特徵，以產生另一新的查詢指令特徵向量加以執行。

41. 如申請專利範圍第 40 項之中文資訊檢索方法，其中該索引特徵之增加或刪除可由該索引特徵常出現於之前檢索所獲得之相關資訊記錄或不相關資訊記錄中加以判斷。

42. 如申請專利範圍第 41 項之中文資訊檢索方法，其中若該索引特徵常出現於之前檢索所獲得之相關資訊記錄中，則增加該索引特徵或其分數。

43. 如申請專利範圍第 41 項之中文資訊檢索方法，其中若該索引特徵常出現於之前檢索所獲得之不相關資訊記錄中，則刪除該索引特徵或降低其分數。

44. 如申請專利範圍第 15 項之中文資訊檢索方法，另包含有於利用該索引特徵檢索欲查詢之以語音或文字型式呈現之資訊記錄之步驟後，進行一第二次檢索。

45. 如申請專利範圍第 44 項之中文資訊檢索方法，其中該第二次檢索可由增加索引特徵或刪除索引特徵，以產生另一新的查詢指令特徵向量加以執行。

46. 如申請專利範圍第 45 項之中文資訊檢索方法，其中該索引特徵之增加或刪除可由該索引特徵常出現於之前檢索所獲得之相關資訊記錄或不相關資訊記錄中加以判斷。

47. 如申請專利範圍第 46 項之中文資訊檢索方法，其中若該索引特徵常出現於之前檢索所獲得之相關資訊記錄中，則增加該索引特徵或其分數。

48. 如申請專利範圍第 46 項之中文資訊檢索方法，其中若該索引特徵常出現於之前檢索所獲得之不相關資訊記錄中，則刪除該索引特徵或降低其分數。

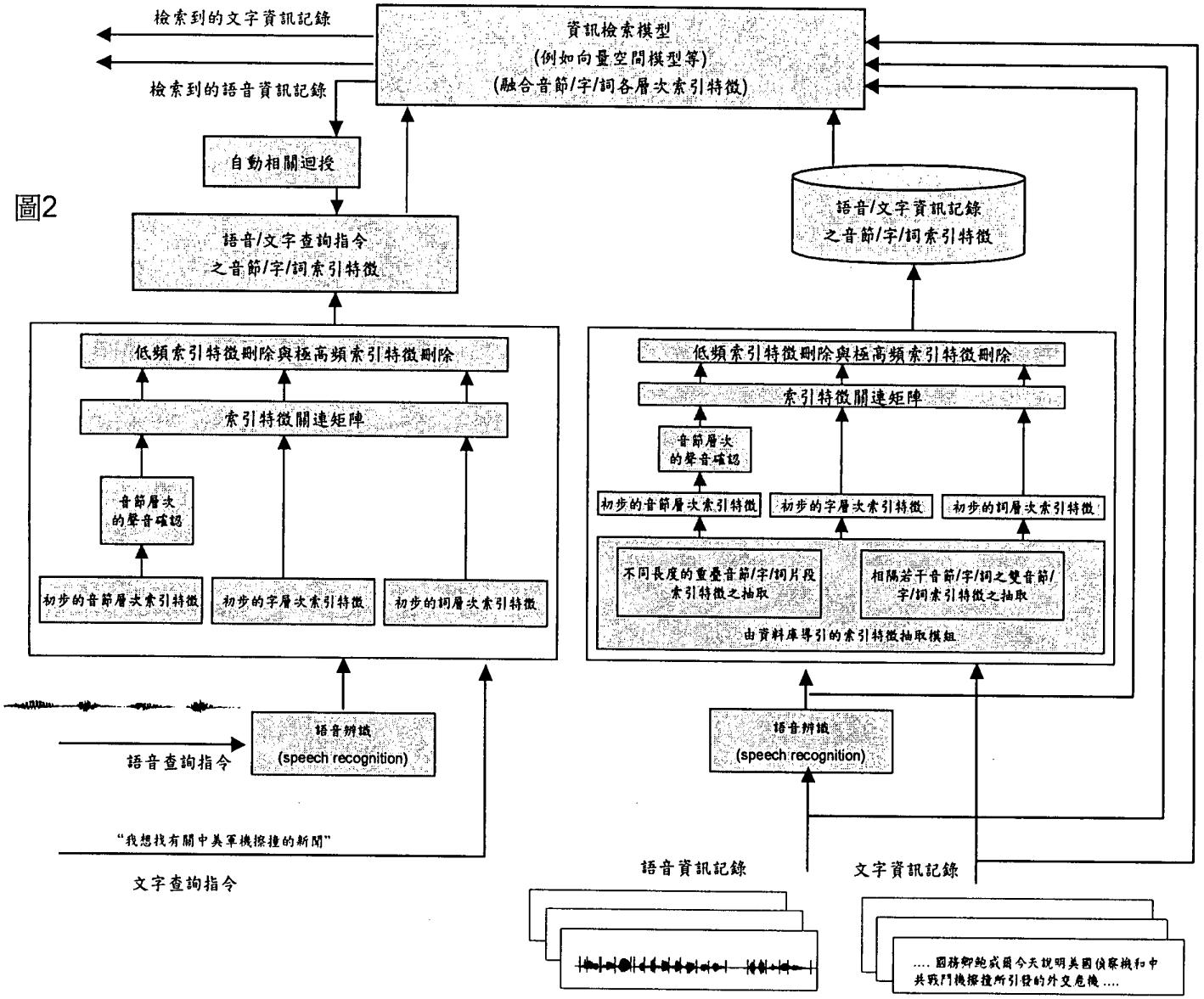
拾壹、圖式

不同長度重疊音節片段 (Overlapping Syllable Segments with Length *N*)	範例
S(N), N=1	$(s_1)\ (s_2)\ \ldots(s_{10})$
S(N), N=2	$(s_1\ s_2)\ (s_2\ s_3)\ldots(s_9\ s_{10})$
S(N), N=3	$(s_1\ s_2\ s_3)\ (s_2\ s_3\ s_4)\ldots(s_8\ s_9\ s_{10})$
S(N), N=4	$(s_1\ s_2\ s_3\ s_4)\ (s_2\ s_3\ s_4\ s_5)\ldots(s_7\ s_8\ s_9\ s_{10})$
S(N), N=5	$(s_1\ s_2\ s_3\ s_4\ s_5)\ (s_2\ s_3\ s_4\ s_5\ s_6)\ldots(s_6\ s_7\ s_8\ s_9\ s_{10})$
間隔若干音節之雙音節 (Syllable Pair Separated by *n* Syllables)	範例
$P_s(n)$, *n=1*	$(s_1\ s_3)\ (s_2\ s_4)\ \ldots(s_8\ s_{10})$
$P_s(n)$, *n=2*	$(s_1\ s_4)\ (s_2\ s_5)\ \ldots(s_7\ s_{10})$
$P_s(n)$, *n=3*	$(s_1\ s_5)\ (s_2\ s_6)\ \ldots(s_6\ s_{10})$
$P_s(n)$, *n=4*	$(s_1\ s_6)\ (s_2\ s_7)\ \ldots(s_5\ s_{10})$

圖 1

圖式

圖2


檢索到的文字資訊記錄
檢索到的語音資訊記錄
資訊檢索模型
(例如向量空間模型等)
(融合音節/字/詞各層次索引特徵)
自動相關迴授
語音/文字查詢指令
之音節/字/詞索引特徵
語音/文字資訊記錄
之音節/字/詞索引特徵
低頻索引特徵刪除與極高頻索引特徵刪除
索引特徵關連矩陣
音節層次
的聲音確認
初步的音節層次索引特徵
初步的字層次索引特徵
初步的詞層次索引特徵
不同長度的重疊音節/字/詞片段
索引特徵之抽取
相隔若干音節/字/詞之雙音節/
字/詞索引特徵之抽取
由資料庫導引的索引特徵抽取模組
語音辨識
(speech recognition)
語音查詢指令
"我想找有關中美軍機擦撞的新聞"
文字查詢指令
語音資訊記錄
文字資訊記錄
.... 國務卿鮑威爾今天說明美國偵察機和中
共戰鬥機擦撞所引發的外交危機 ....